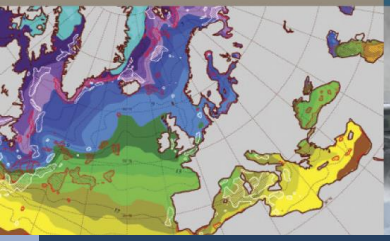
ECMWF

GLOBAL PREDICTION | SEVERE WEATHER | ATMOSPHERIC COMPOSITION | CLIMATE MONITORING | SUPERCOMPUTER CENTRE





When a Computer Hall Starts to Crack at the Seams

Moving the ECMWF Archive

September 2, 2016

European cooperation at its best

ECMWF's role is to address the critical and most difficult research problems in medium-range NWP that no one country could tackle on its own

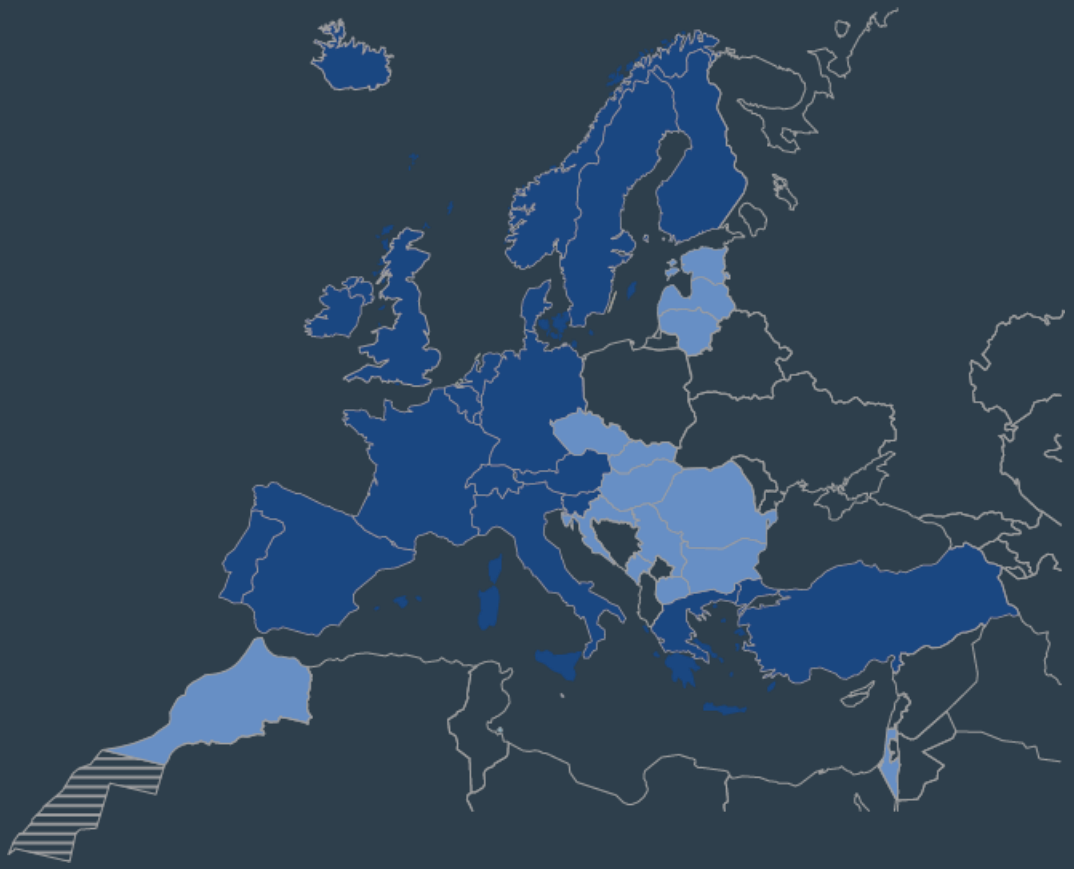
European cooperation at its best:
Deliverables and research

- Global numerical weather forecasts
- Composition of the atmosphere: monitoring and forecasting
- Climate reanalysis: monitoring
- Supercomputing & data archiving
- Education programme

European with a global reach

- 34 member and co-operating states
- 270 staff
- 30 countries
- Partnerships around the world …

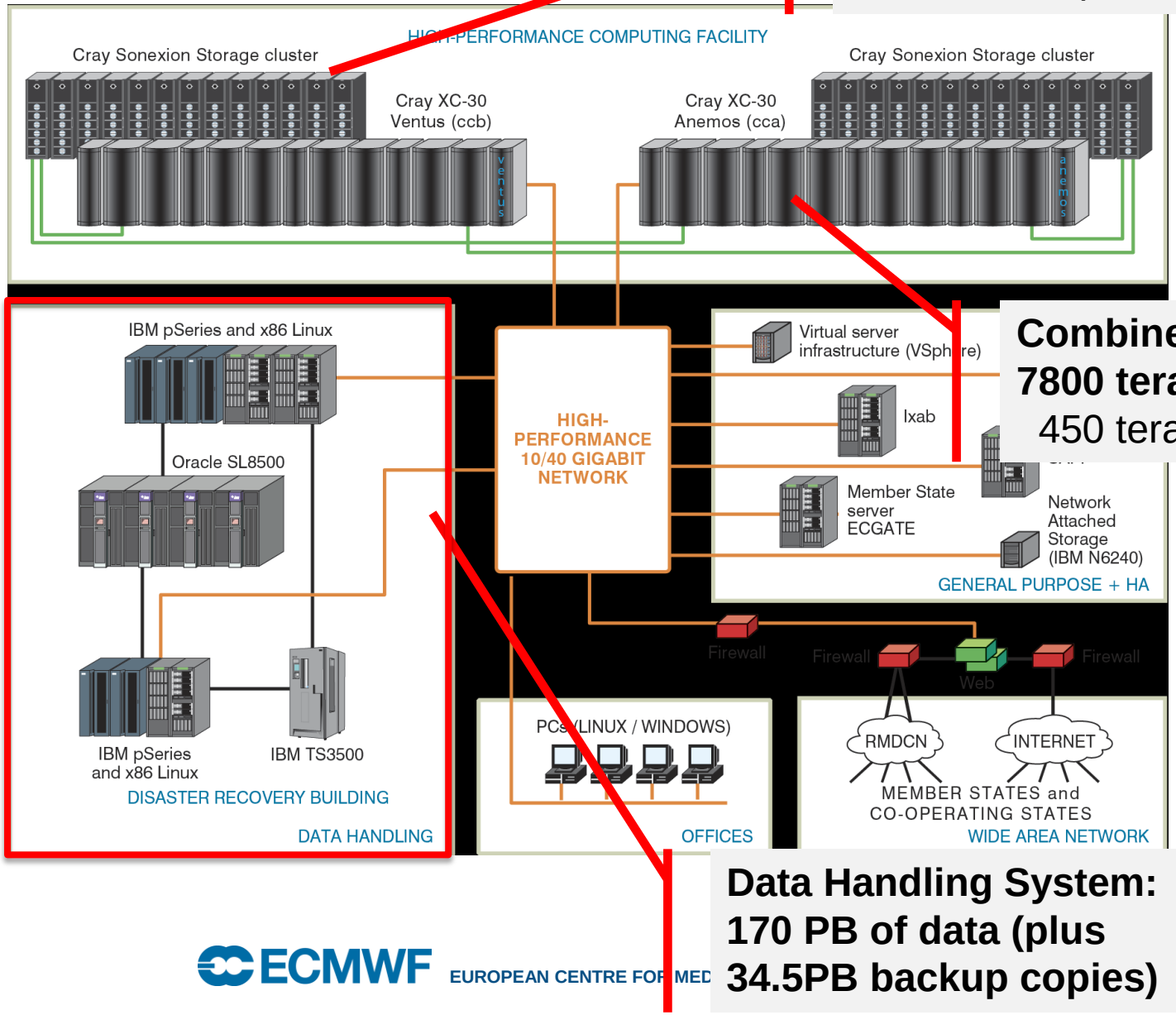
International cooperation at its best



ECMWF

Our computing centre

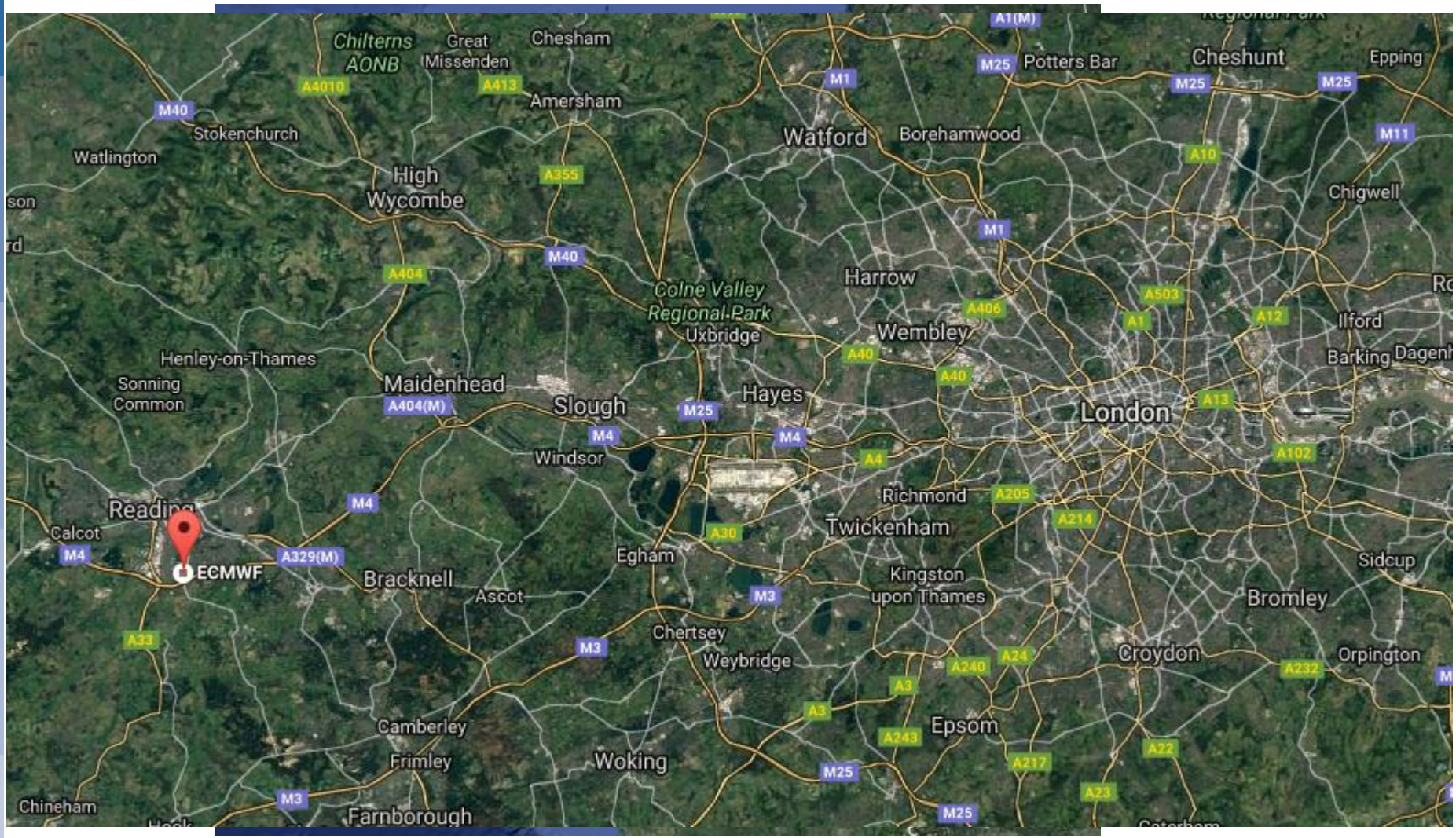
Lustre clusters:
About 20PB (combined)



Combined Power:
7800 teraflops (peak),
450 teraflops (sustained)

Data Handling System:
170 PB of data (plus
34.5PB backup copies)

Where is ECMWF?



ECMWF Shinfield site.

- HQ
- Includes
  - Offices for most staff,
  - Computer halls.
- Drawbacks:
  - 40 years old
  - Limited Space
  - Limited power supply
  - Limited cooling facilities

ECMWF

ECMWF Reading Enterprise Centre

- Additional office space at the University of Reading.
  - Purely for research staff supporting an EU founded project
- No significant local computing facilities
- Access to computing facilities via leased fibre lines.
- 6 Miles away.

2020: New High Performance Computer

- A new set of HPC clusters to be installed
  - Roughly 2.5 times more powerful than existing machines.

- The problems:
  - Where to get additional power and cooling?
  - Where to fit this new machine in the limited space available?

- Not cost efficient to do this on the existing site.

Time to move!

The high-level solution

- Let's build new facilities!
  - Sufficient to accommodate the growing number of staff
  - Sufficient to support the 2020's HPC
  - That can be extended to support larger HPCs in the future.
- To be ready by 2H2019,
  - New HPCs to be built on that site in early 2020
- DHS, and other computing equipment to migrate there.

	Power	Computer hall area	
Currently	4.2 MW	2,250 m^2	24,200 ft^2
2020	Up to 10 MW	3,000 m^2	32,300 ft^2
Could be extended to	Up to 20 MW	5,000 m^2	53,900 ft^2

A difficult decision

- Discussions about replacement site ongoing for many years
- Any solution proposed needs approval from ECMWF's Council.
  - Representatives of governments from 21 Member States.
  - Meeting twice a year.

- A number of iterations took places, over the years, trying to define
  - Where to move
  - What to move
  - Whether to introduce geo-redundant solutions.

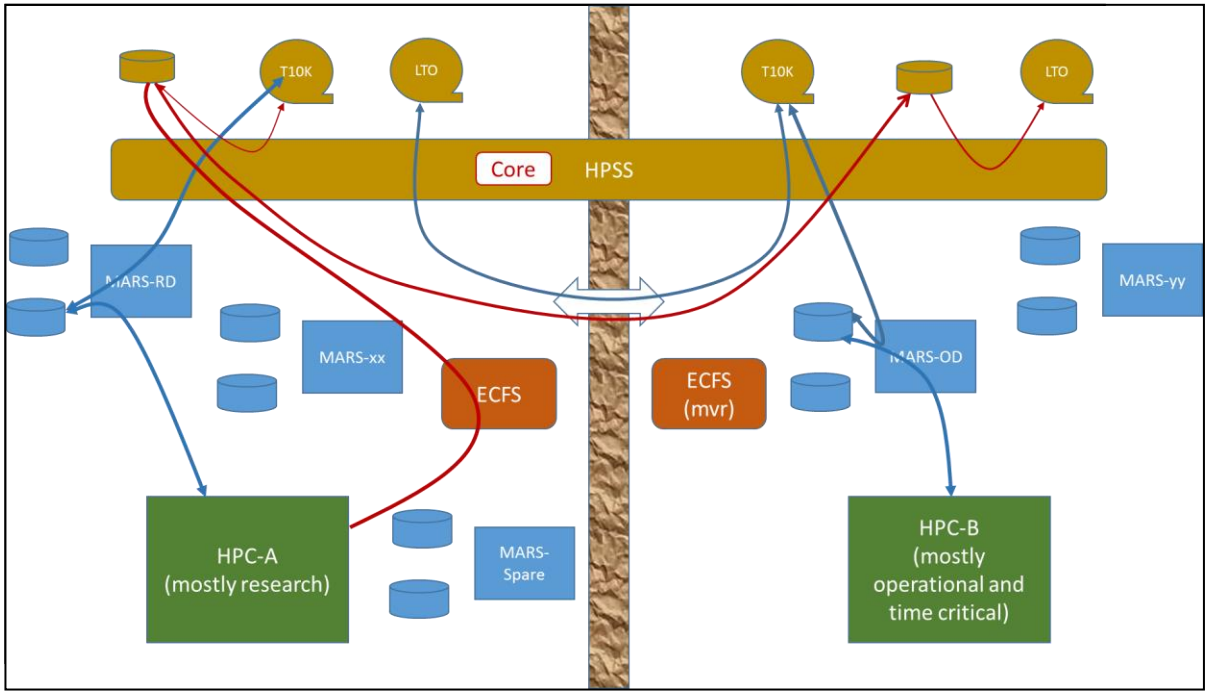
- Decision possibly in December 2016?
  - Only 2 years and a half to built new facilities, prepare migration, etc.

Computer Halls Scenarios Considered.

- Relocation of all ECMWF Computer halls to one new site
  - New HPC installed on a new site
  - DHS and most other servers are moved to this new site
  - Some small servers (e.g. nfs) may be kept with the offices
- Relocation towards two (geo-redundant) Computer halls
  - First HPC cluster moved to one new site,
  - Second HPC cluster moved to another new site
  - DHS
    - moved to one of these new sites,
    - Split between two new sites
- Partial relocation (keeping Shinfield's Computer hall in operation, but installing some of the computing environment on another site)
  - Variation on the previous theme.

Dual sites options

- Distribute HPC and DHS across the sites
  - Better resilience to catastrophic event
  - Critical data and one HPC on each site allow to switch operational service.
  - Each site would favour some type of non operational work, to limit WAN traffic.



Dual sites issues

- Costs of WAN connections expected to be considerable.
- Other issues
  - To reduce traffic between sites:
    - Keep types of non-operational suites to specific sites
      - Loss of flexibility, under-utilisation of HPC resources.
  - If sites are far away,
    - More duplication of operational data on HPC file systems or
    - Much more reliance on DHS when switching operation from HPC to HPC.

- This option was abandoned.
  - For the time being…

The DHS migration challenge

- Avoid operational work outages due to the relocation.
  - Limit Research/other work outages to a minimum.
- Support new HPCs before they go operational.
  - HPC acceptance.
  - For 6 months or more.
- Access to the archive by operational HPC clusters needed 24x7.
  - In 2020, bandwidth > 100Gib/s just to support day to day activity.
- Short Shinfield lifetime once the Operational models are transferred.
  - Computing halls to be released by end 2020. 5 months or less..

Keep costs to the bare minimum

Copying the data across the network ?

- “Great time to change technology if needed”.
- By mid 2020, at least 510 PB in Shinfield.
- Migration time if one migration stream only:

30,000 days
(1,000 months)

Assumes 200MB/s sustained (not necessarily realistic)

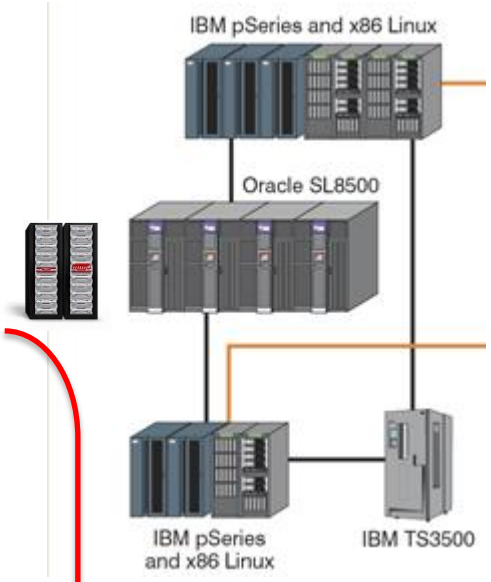
- To complete in 5 months:
  - Number of parallel streams: **200 streams**
  - Number of drives: **420 enterprise-quality tape drives**
  - Inter sites bandwidth: **320 Gib/s links**

Not really affordable.

Data – on tapes – will be transferred by the truckload

What about equipment?

- Tape libraries (Primary)
  - Moving these would take up to four weeks.
  - Will need new libraries on new site.
- T10K Primary tape drives move with the tapes.
- Tape libraries (backups)
  - Could be moved in a week
- Disk systems (16PB)
  - Too expensive to duplicate or replace.
  - Could be moved fairly quickly
  - Risk factor ➔ all “important” data copied to tape 1st.
- Servers, LAN, SAN
  - Buy minimum of new kits for the new site ahead of time.
  - Complement later by equipment coming from Shinfield.



Up to two weeks to move

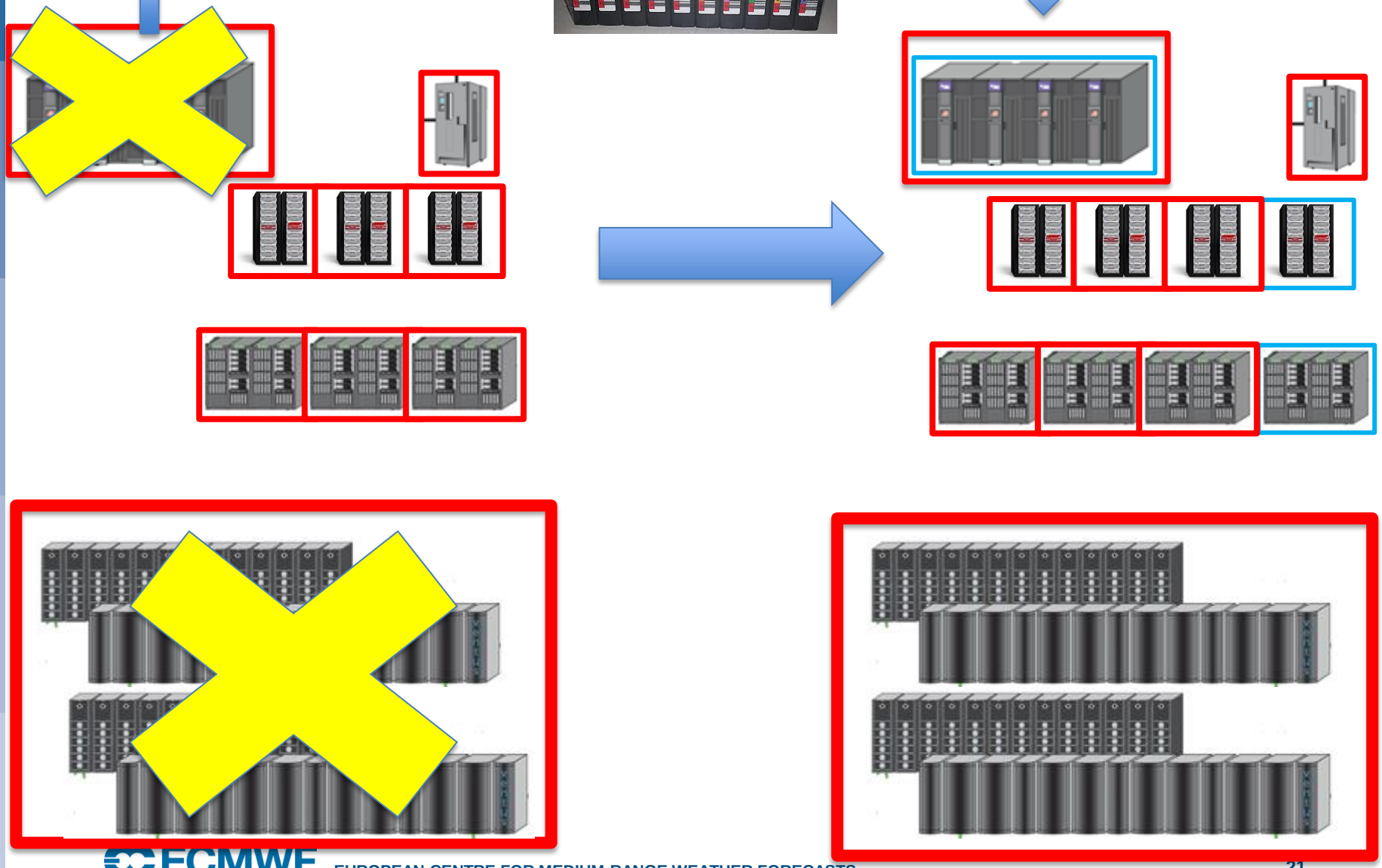
How do we provide continuity of operational service?

- Moving and redeploying DHS equipment will take at least two weeks,
- Tens of PB of operational data generated during that period.
- HPC need to be able to retrieve information to run Operational models
  - Especially if HPC file systems are hit by issues.

How do we provide service to new HPC during acceptance?

- Network bandwidth between sites will be limited
  - No major investment to cover a short transition
- New site will have to store and retrieve most data locally.

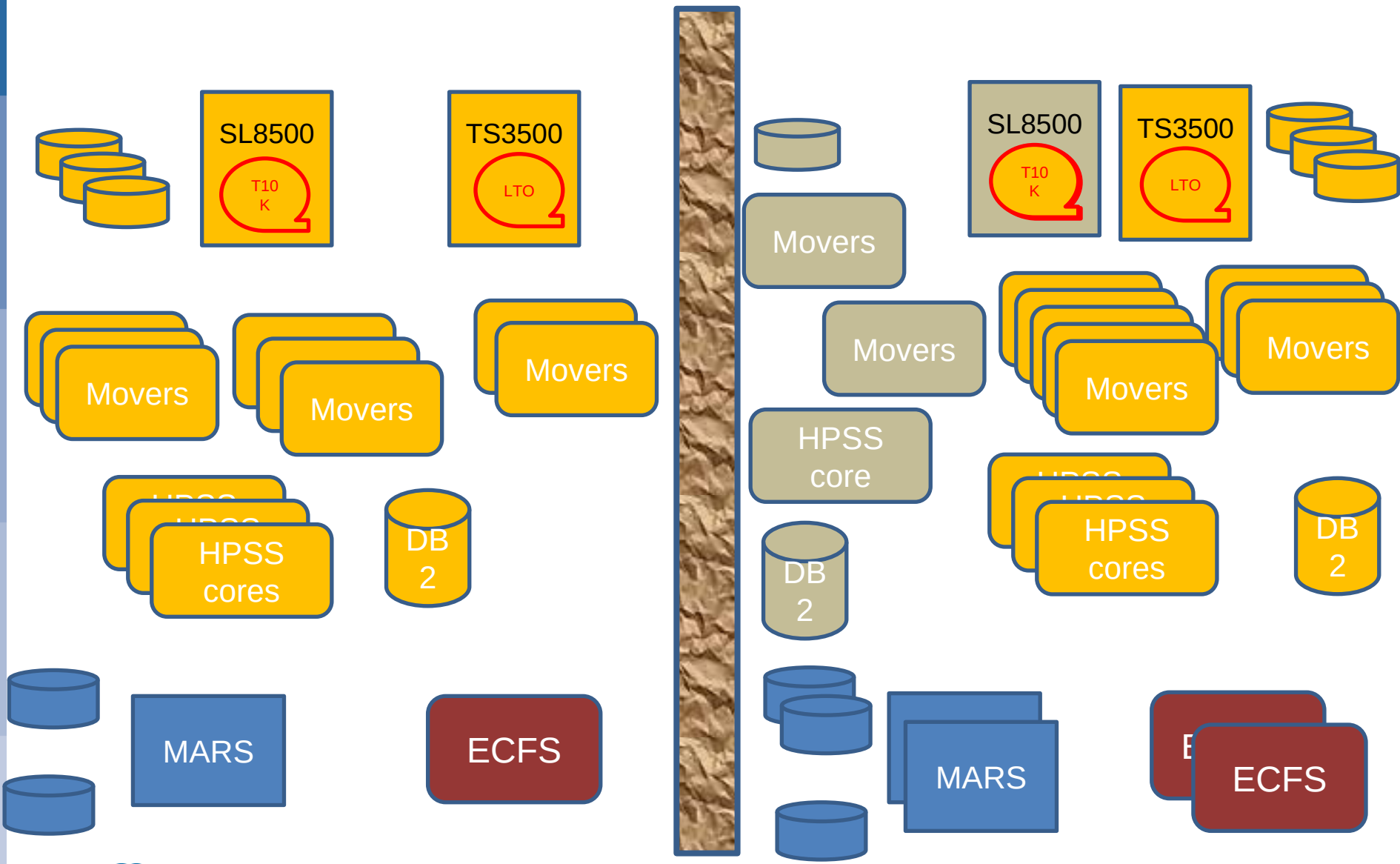
Migration overview



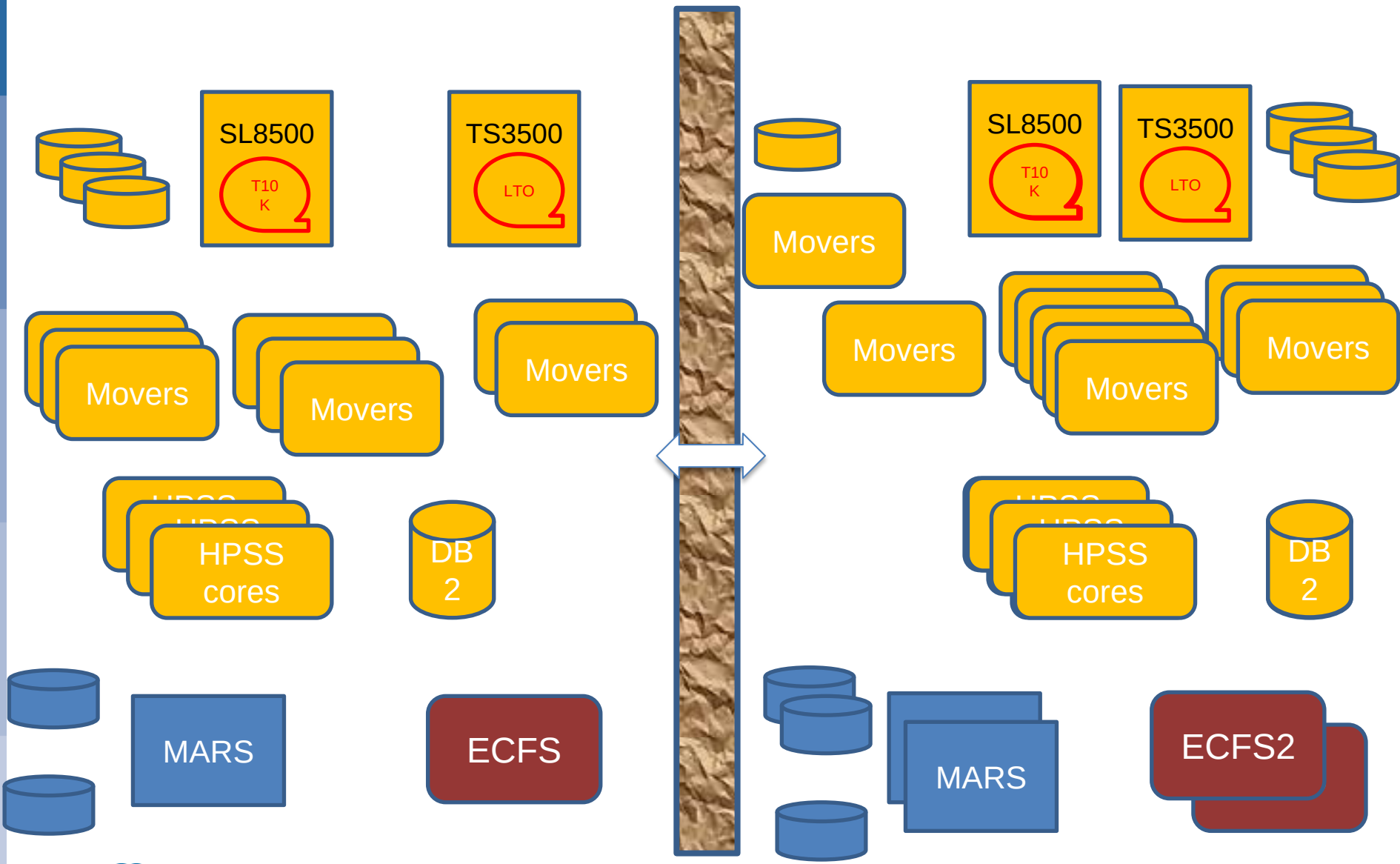
What does this means in HPSS terms?

- Two options considered:
  - Use of a second HPSS instance
  - Distributed HPSS environment.

Option 1: Use a second HPSS instance.



Option 2: distributed HPSS



Pros and Cons of both solutions

Second HPSS instance

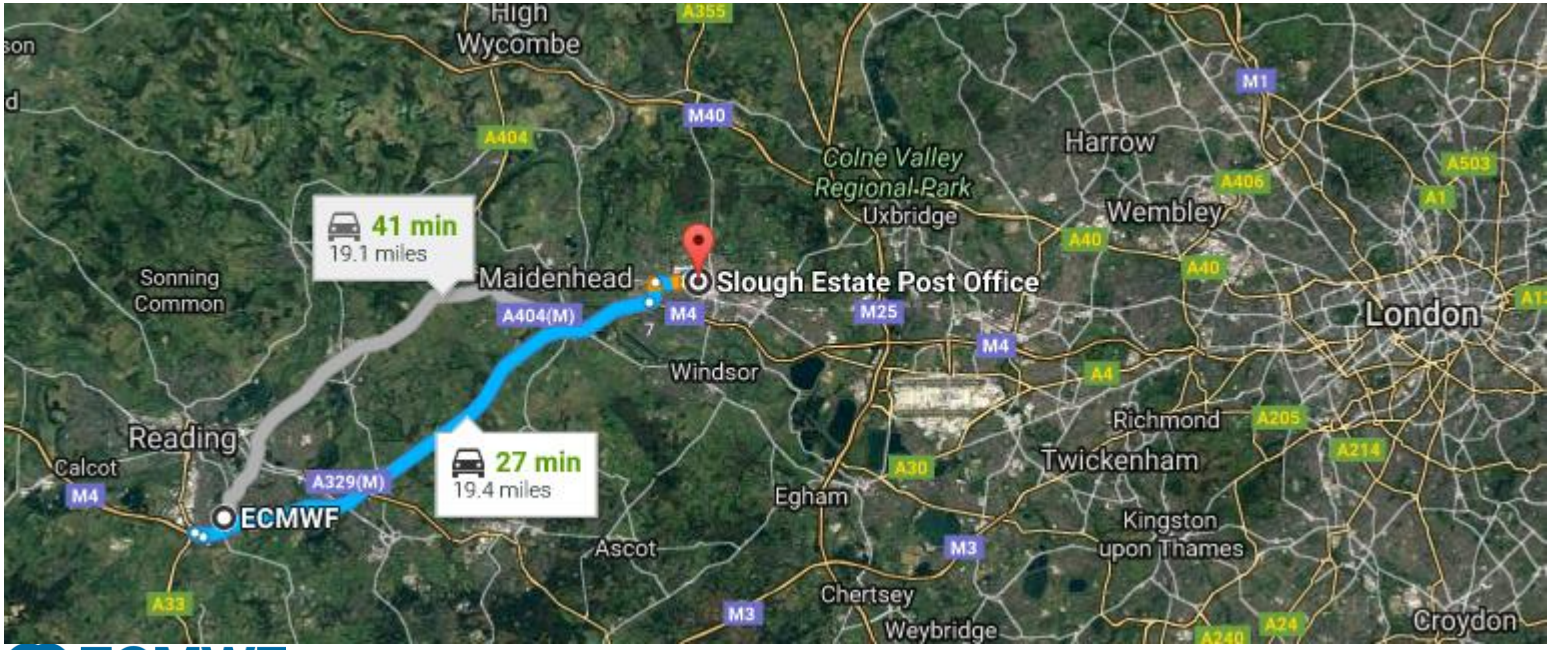
- Less dangerous operational transition
- No dependency on network
- Strict separation between sites
- Needs complete HPSS setup
  - Require second licence
- No access to some data for several weeks
- Applications:
  - Where is the data stored?
  - Copy the data in main instance

Distributed HPSS environment

- Switching cores: “Big bang”
  - May affect operations
- Vulnerable to network problems
- Limited cross-access possible
- Needs separate SCs, Hiers. COS.
  - But all data in same name space
- Access to most data can be restored quickly.
- Applications:
  - Which family to use?
  - Copy data back in normal hierarchies

Computer Halls vs offices.

- A possible scenario:
  - Computer halls are built in fairly close proximity (Slough, next to Windsor)
  - Offices stay In Reading
  - 30KM (20M) away.
    - On a very busy Motorway…



What if the halls are moved a bit further away?

- HQ and offices stay in the Reading area
- The Computing Facilities could move to Iceland
- A mere 1800 KM (1100 Miles)
  - A bit far away to provide hands on support…
- Any experience to share?

In Summary

- ECMWF will not be able to install new HPCs on its current site.

- New site expected to be put in service in 2020.
  - Still don’t know where.
- Data Handling System (including HPSS) will be transferred to new site.
  - Challenging
  - Transition to new site must not impact operational work.
  - Will require a limited duplication of DHS resources.

- Management of DHS expected to be done remotely.

Francis.Dequenne@ecmwf.int

Ian.Randall@ecmwf.int